AD-A136 275 FUNCTIONAL AND DATABASE ARCHITECTURE DESIGN(U) ALPHA OMEGA GROUP INC HARVARD MA 26 SEP 83 N00014-83-C-0525

1/1

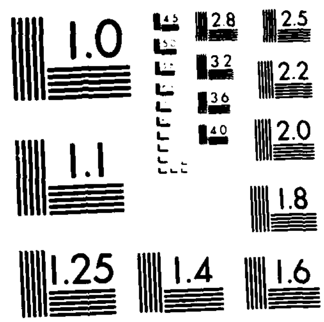
UNCLASSIFIED F/G 5/2 NL

END

DATE FILMED

1-84

DTIC



MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS 1963 A

A136275

(12)

Contract No. N00014 83-C-0525

REPORT A001

FUNCTIONAL AND DATABASE ARCHITECTURE DESIGN

Submitted to:

Office of Naval Research
Department of the Navy
800 N. Quincy Street
Arlington, VA 22217

By:
Alpha Omega Group, Inc.
P. O. Drawer M
Harvard, MA 01451

September 26, 1983

alpha omega group, inc.

TABLE OF CONTENTS

REPORT ON

FUNCTIONAL AND DATABASE ARCHITECTURE DESIGN

1. STATEMENT OF THE PROBLEM

Much attention and considerable effort has been focused recently on the collection and storage of descriptions of an organization's data and information resources. Some organizations have been quite successful in this collection and storage effort. A different problem, which is more difficult to deal with, has been to provide an easy, effective mechanism for users to access this information once it is stored. This problem generally reduces to determining the existence of information.

The determination of the existence of required information is a non-trivial problem; there are a number of possible answers. These include:

- The information exists in complete and appropriate form.
- The information exists, but is not in a form which is appropriate based on the user request.
- Some, but not all, of the information exists.
- All of the data exists, but it must be synthesized to produce the desired information.
- Some data exists which may be synthesized to produce some of the desired information.
- The information exists, but is in a larger set or context. This is a case where a request has been made which is too specific for the information breakdown available. An example might be a request for census information provided at the county level, where information is available only on a state-wide basis.
- The information does not exist as requested, although alternative information may exist which could be relevant to the user's needs.

In addition to the simple existence issue, there are also qualitative issues which apply to information requests. These include timeliness, appropriateness, accuracy (i.e., precision), and the like. Data Dictionary Systems (DDSs) have proven to be reasonably successful in documenting these qualitative criteria, but the existence issue is one which is still largely ignored by the information system community.

The Locator and Classifier for Universe Standardization (LOCUS) is a concept which seeks to provide a tool which will aid the user in determining the existence and location of the "information about data" (i.e., metadata) which is required to perform some task. It is important to emphasize here that LOCUS is a system which operates on metadata, not data itself.

Currently, most dictionary systems provide some assistance in solving the problem stated above by providing an ability to associate installation-standard keywords with dictionary entities, and query facilities to identify entities based on any Boolean combination of those keywords. While helpful in locating entities, these facilities have several limitations:

- Associating keywords with entities is a purely manual process. The software more often than not provides no assistance in assuring that associated keywords are correct and adequately characterize the entity.

- The list of appropriate keywords has often been developed in an ad-hoc fashion. There is no underlying discipline, methodology, or approach to classifying entities.

- Keyword lists typically classify entities at one level of abstraction. These tend to be too general.

The effects of these limitations are as follows:

- Entities may be incompletely characterized. Thus they may be missed in searches, and may even exist redundantly.

- Searches may be frustrating because the categories for selection may be too broad. Once a list is given, "zooming in" on the desired entities is often done entirely by scanning the list provided.

2. LOCUS USER INTERACTION

The goal of LOCUS is to provide a very friendly interface for users who have no idea of the name by which the entity is known in the dictionary, but who know certain characteristics of this entity or related entities.

Before describing LOCUS user interaction, it is first necessary to characterize the LOCUS user. Let us assume that the user belongs to an organization whose dictionary is being used to locate information (e.g., an information resource manager, a bank employee, a secretary, a supervisor, etc.). This user will have:

- Specialized knowledge of his/her particular function within the organization.

- At least a general knowledge of the functions of the organization (i.e., the goods and or services provided by the organization.)

- Knowledge of the general organizational structure.

- Knowledge of likely combinations of data types used by the organization within his particular job function.

- Knowledge of input and output and standard forms used by the organization.

- Basic concepts of space and time.

- Knowledge of basic data classes (e.g., date, time period, amount, etc.)

LOCUS is composed of two major phases. These are:

- The selection phase, and

- The display phase.

The selection phase uses what we will later define as aspect trees and selected dictionary entity descriptions to select the desired dictionary entities. The display phase provides the user with the detailed information on the entities that have been selected in the first phase. This document addresses only the first phase.

The selection phase will operate in one of two modes. The first mode will provide user prompting, resulting in a type of dialogue to assist the user in finding the desired information. The second mode will make the assumption that the user has knowledge of the classification aspects within LOCUS and is therefore able to specify the appropriate combinations in a command sequence. Facilities will be provided to allow the user to move from one mode to the other. In this manner maximum flexibility can be provided to the user.

3. DISCUSSION OF LOCUS CONCEPTS

Following are discussions of basic concepts which are necessary in providing a sound understanding of the LOCUS system. The concepts to be introduced are:

- Data Dictionary Systems,

- Classification Theory,

- The Thesaurus Facility, and

- LOCUS Searching and "Hits".

3.1 Data Dictionary Systems

The concept of a Data Dictionary System originated from the need for a centralized repository for storing definitions and descriptions of an organization's data. This "information about data" is known as metadata, and metadata is an essential component in the effective management and control of the data and information environment. Proper use of this metadata is required to ensure consistent data documentation and to control data access and usage.

The proliferation of Data Dictionary Systems throughout the Federal Government has prompted the Institute of Computer Science and Technology of the National Bureau of Standards to initiate a standardization project leading to the specification of a Federal Information Processing Standard (FIPS) for a Data Dictionary System (DDS). Since both a Data Dictionary System and LOCUS deal with metadata, we assume that the facilities of a Data Dictionary System are available in the LOCUS environment. The existence at this time of the FIPS DDS specification and its importance in the federal sector leads us to make the further assumption that the LOCUS design should be placed in the context of the FIPS DDS. Variants of this design can be developed which are applicable to other Data Dictionary Systems.

When first conceived, DDSs were limited to the description of those entities and relationships which relate only to data. Typical DDSs included references to the following types of entities (i.e., "entity-types"):

ELEMENT,
RECORD,
FILE, and perhaps,
DATABASE.

Some of these Data Dictionary Systems later began to include additional entity-types such as:

DOCUMENT,
FORM, and
REPORT.

As DDSs evolved, their scope expanded to include not only descriptions of the data in an organization, but also descriptions of processes which take place within the organization. Examples of such entity-types are:

TASK,
PROCEDURE,
PROGRAM,
MODULE,
SYSTEM,

and the like. Modern-day DDSs have gone beyond even these entity-types, and are being designed to encompass descriptions of the entire range of information resources of an organization. These systems will typically include additional entity-types such as:

EQUIPMENT,
WORKSTATION, and
USER.

By an entity of type USER, we do not mean people who use the DDS, but rather organizational components or roles.

Some of these entity-types are "simple" entity-types, meaning that they represent atomic entities which are not made up of other entities. Other entity-types, such as DOCUMENT and REPORT, are "compound" entity-types which can be viewed as "containers" for other entities.

The basis of the DDS is the dictionary schema, which describes the structure of the dictionary, and the constraints and rules to which the metadata must conform. The schema will include, for example, the entity-types, relationship-types, and attribute-types necessary to describe the information environment. In addition, the dictionary schema may include the structure for metadata security constraints, metadata validation rules, system life cycle support, and the like.

3.2 Classification Concepts

The final report for Phase I of this contract provided an approach for describing complex topics which made use of classification theory. Classification schemes may range from simple hierarchies of all conceivable topics in the domain to be classified, to so-called "analytico-synthetic" schemes by which complex topics can be analyzed into their constituent aspects, and a full description synthesized from those aspects so that this description (or any subset of it) can be searched for and located.

3.2.1 Aspect Classification

"Faceted classification" is a special kind of analytico-synthetic classification which draws component parts from specially-constructed lists (called facets) which derive from the application of single, specific characteristics. This approach is based on the definition of facets in terms of the following fundamental categories:

- THINGS,
- PROPERTIES of the THINGS,
- ACTIVITIES, FUNCTIONS or PROCESSES involving THINGS,
- TOOLS that support the ACTIVITIES,
- human or institutional PARTICIPANTS in the ACTIVITIES, and
- indications of LOCATIONS and TIME.

Faceted classification exists as a free-standing facility in the Library and Information Science field, while LOCUS assumes the existence and availability of a DDS. This DDS provides, in itself, a substantial amount of information which can be made available to the user, including relationships between entities. The approach used in this document is based on a variation of the faceted approach. Faceted classification has been shown to be valuable in the classification of elements; however, it is not readily adaptable to classifying entities of other types in the information environment. Hence, the requirements on the classification scheme in this environment will be different from those addressed by the faceted classification scheme. In the LOCUS environment, we will use the term "aspect" in a fashion similar to that of "facet" in the faceted classification environment.

Following are definitions related to the LOCUS classification scheme. They are presented to support the discussion which follows in this report.

- ASPECT NAME - In defining LOCUS, we will assume the existence of a set of classification aspects, each characterized by a unique "aspect name".

- ASPECT TREE - With each aspect name there is associated an "aspect tree", which is a hierarchical structure made up of "normalized terms" (defined in the next section) which are associated with the classification aspect. The root of each aspect tree is the aspect name.

- KEYWORD - At each node of these aspect trees is a normalized term which we will call a "keyword".

- ENTITIES - The objects which are classified will be referred to as "entities". Keywords are associated with these entities.

An aspect may apply to only one, to many, or to all entity-types.

3.2.2 Normalized Language

Keywords, as defined above, are the vocabulary of the LOCUS normalized language. By a "normalized language" we mean the following:

- a standard set of terms (keywords) with precise meanings;

- a limited standard set of grammatical constructs to combine those terms.

The purpose of a normalized language is to reduce expressions to a form which will consistently reveal equivalent concepts. Thus, two different expressions may be compared. Use of a normalized language thus facilitates the identification of duplicates and assists searching for desired entities.

The concept of normalized grammatical constructs can best be clarified by an example. Consider:

John loves Mary.
Mary is loved by John.

Both sentences have identical meaning, but the second sentence uses the passive form. A rule of a normalized language for declarative sentences might be to always use the active form, i.e., the first sentence above.

3.2.3 The "OF" Language

The "OF" language is a crude example of a normalized language, which is currently used in DDSs to identify and classify data elements. In the "OF" language, all elements are identified as belonging to precisely one of the following classes:

- ADDRESS - An instance of the element is a geopolitical address.
- AMOUNT - An instance of the element is an amount of money.
- CODE - An instance of the. element is a code.
- CONSTANT - An instance of the element is a constant.
- CONTROL - An instance of the element is a value that is used to control the flow of processes.
- DATE - An instance of the element is a date.
- DESCRIPTION - An instance of the element is a text string which is used as a description.
- NAME - An instance of the element is a name of a person or company.
- NUMBER - An instance of the element is a number.
- PERCENT - An instance of the element is a ratio expressed as a percentage.
- QUANTITY - An instance of the element is a number representing a quantity (including fractions) of anything excepting money; a unit of measure is implicitly associated with this number.
- TIME-PERIOD - An instance of the element is an interval of time; a unit of measure is implicitly associated with this interval.

These terms are defined so there is no overlap and that all data elements belong to one of these classes.

The "OF" language has been demonstrated to be useful in many applications. However, it does have several limitations:

- It applies only to elements.
- It does not easily handle more complex concepts.
- It accomodates only a single level of qualification.

It would be desirable to extend this facility to describe other types of entities, such as reports or documents. Consider, for example, a report on "Census Data on Family Income by Geographic Area". To truly describe or classify this report using only an "OF" language would be impossible. It is necessary, therefore, to provide a more powerful facility, such as the aspect concept, although some aspects may be based on the "OF" language.

3.3 The Thesaurus Facility

The assumption is made throughout this document that a Thesaurus Facility will exist as a component of the LOCUS architecture. The concept of a normalized language is helpful from the perspective of system, but may be very unfriendly from the users' perspective. The thesaurus will provide a mechanism for translating user terms into the normalized terms used by LOCUS.

The contents of the thesaurus should be expandable. Whenever a new user term arises which is not in the thesaurus, facilities will exist in LOCUS to add this term to the thesaurus along with the required references to the normalized language.

3.4 LOCUS Searching and "Hits"

A normalized vocabulary addresses only part of the problem expressed in Section 1. In searching for types of information or data, we recognize that the information or data may exist in many different forms and be derivable from many different sources.

When a LOCUS user is searching for information or data, the following will occur:

- o The LOCUS user request will be reduced to a set of normalized keywords. These will be referred to as query keywords.

- o A query keyword will produce a hit on an entity if there exists a keyword associated with the entity which is either:

  - equal to the query keyword, or

  - a descendant (of the query keyword) in the aspect tree.

Depending on how one formulates a request, the request may result in no "hits" at all. For example, suppose we want to look at archives managed by the National Weather Service. Now, if we were interested in obtaining reports on snowfall in St. Joseph, Michigan, we might try searching based on Snowfall and St. Joseph, Michigan. (NOTE: At this point we are interested in determining the existence of snowfall information, not snowfall amounts.)

No hits may occur because of the many potential ways of combining the weather information. We could have Weather/National, Weather/State, Precipitation/City, etc. In order to deal with this we must apply the concepts of classification. We have here two distinct concepts: Geography and Weather. Each of these can be represented by an aspect tree as follows:

```
                           Weather
                              |
     +-----------+------------+-----------+-----------+
     |           |            |           |           |
Temperature    Wind        Precipi-    Humidity     Cloud
             Velocity       tation                  Cover
                              |
                 +------------+-----------+
                 |            |           |
                Rain        Sleet        Snow
```

and

Geography
|
Worldwide
|
Continental
|
National
|
State
|
City

Although we missed getting an exact hit based on a breakdown of "Snow" and "City", this does not mean the information does not exist. Since cities are parts of states, and snowfall is a form of precipitation, we might reasonably expect to find precipitation records collected by state, precipitation by city, or snow by state. Each one of these combinations is "close to" the combination which we sought. We would expect to find the detailed ir rmation we want as line items in records or reports belonging to one of thes ombinations.

Thus, we see that in searching for types of information or data, fin g the "near miss" may be more important than the exact "hit", simply becau near misses are much more likely.

We also see that aspect trees give us a framework to define even what we mean by "near misses".

4. LOCUS USAGE SCENARIOS

To illustrate the need for LOCUS and the LOCUS interface, several scenarios have been developed. These scenarios address potential problems which exist in most information environments, whose solution could be assisted by integrating a classification approach with a DDS.

4.1 Scenario 1 - Information Retrieval

Many staff personnel at a headquarters or, for example, on a congressional committee, are tasked to find current information about a subject area. The common approach either involves the existence of corporate knowledge concerning this information (or potential sources of this information), or to task someone to collect or generate the information because there is no way to locate it among the information systems of the organization.

Resolution of this problem first requires identification of the elements which correspond either directly or potentially to the information sought. To correspond directly implies that the element is the piece of information or it can be derived via some calculation or decision. To correspond potentially implies that source data exists which may be used to derive the desired information. This first part can be satisfied by a classification scheme.

The second part of this problem is to locate the container which holds the desired information and/or to locate the processes which are used to generate the information. This is a natural application for the DDS and the relationships which are described in it.

4.1.1 Implications of Scenario 1

In addressing the implications of Scenario 1, it is assumed that the DDS is populated with the necessary elements, containers, and processes, in addition to the relationships between the entities.

Thus, the requirements resulting from Scenario 1 are as follows:

- The user would provide terms, which for now we will assume are normalized. These terms, which correspond to nodes in aspect trees, could be used to identify elements which have:

  - all of the terms as keywords; this would imply an "AND"ing operation and results in a direct hit. At this point the user could be provided the names of the entities and the opportunity to get more information about each entity, such as its description. Also, the user could identify related entities, based on relationship-types. Related entities might be directly or indirectly related. It should be assumed that the user only needs to provide an entity-type and the system would find any related entities.

  - one or more of the terms as keywords; this would imply an "OR"ing operation and results in a potential hit. In this case, the user interface needs to allow the user to either modify his "selection set" of terms to allow identification by using either a more general (a node closer to the root) or a more specific (a node farther from the root) term which is related to one of the previously specified terms. At some point it will be necessary to switch to the "AND"ing approach described above, and perhaps back to the "OR"ing approach.

- If we assume the terms are not normalized, this reasonable problem becomes more complex, because there is a need for a Thesaurus Facility that translates the user's terminology into the normalized form.

4.2 Scenario 2 - Requirements Analysis Support

An analyst has begun a requirements analysis. The purpose of this analysis is to study the existing information environment to find out whether stated requirements can be satisfied by current information systems or modifications to existing information systems, or if they require the development of new information systems. In current environments, even with a DDS, this problem is far from trivial. Even if all the components of the information environment have been documented in the DDS, identification of the "new" requirements in terms of the existing documented environment may not result in hits, even though in reality it should. A classification scheme which allows the precise identification of existing information and its sources would be highly desirable. Although the information sources can be identified through the DDS relationships, the keywords assigned to the Element entities are the potential solution to this problem. Of course, assignment of the keywords must be based on a classification approach if we are to be assured of precise identification.

4.2.1 Implications of Scenario 2

As mentioned in the discussion, to be truly useful this environment must be more complex. The reason for its complexity is that to be useful, a Thesaurus Facility must exist.

4.3 Scenario 3 - Naming Conventions for Data Administration

A data administration function is trying to establish naming conventions, and is experiencing problems identifying and resolving homonym/synonym inconsistencies in entity names. Resolution of this problem implies the ability to associate semantics with, at least, Elements, and to provide software to analyze these semantics. An approach to solving this problem is to provide a controlled vocabulary and an associated classification scheme which uses the controlled vocabulary.

4.3.1 Implications of Scenario 3

Although it would be extremely helpful to have a thesaurus facility, it is reasonable to assume that the data administration staff is assigning normalized keywords to entities, since this is a basic restriction to the vocabulary. The critical part of this problem is the classification methodology, because the methodology should provide a rational approach to assigning keywords to entities. The assignment rules should have the following characteristics:

- The keywords can come from any tree to assure that no redundant terminology is identified, simply because a term is not "known" for the entity-type.
- The keyword should be as far away from the root as necessary to gain precision. A restriction on assigning keywords is that a tree should be "represented" in the keyword list for the entity by only one keyword.
- The assigned keywords should allow unique identification of the Element. This could be accomplished by allowing a special set of keywords which may not be enforced by the classification methodology and which, in fact, are not known to the classification software for either retrieval or assignment. They should, however, be part of the "controlled" vocabulary to allow the data administration function to keep a handle on the problem.

5. LOCUS/DDS ARCHITECTURE AND INTEGRATION

5.1 Components of the Architecture

Figure 1 depicts a simplified view of the LOCUS architecture, which is composed of the following components:

- the User Interface,
- the DDS Management Facility,
- the LOCUS Management Facility,
- the Controller,
- the Maintenance/Reporting Facility,
- the Metadata Database,
- the Meta-Metadata Database, and
- the LOCUS Thesaurus Facility.

These components are defined in the following subsections.

5.1.1 The User Interface

The User Interface is the common user interface to both the DDS Management and LOCUS Management Facilities. The User Interface directs user requests to the appropriate management process. The design of this interface is dependent upon the resolution of a variety of issues concerning the user and user interaction.

5.1.2 The DDS Management Facility

The DDS Management Facility is responsible for translating user inputs into the appropriate DDS maintenance or reporting capability available.

5.1.3 The LOCUS Management Facility

The LOCUS Management Facility controls user interaction with the LOCUS database and schema. This interaction will guide the user in determining the desired information.

5.1.4 The Controller

The Controller acts as the principal interaction control mechanism between the DDS Management and LOCUS Management Facility. It assures that the processes

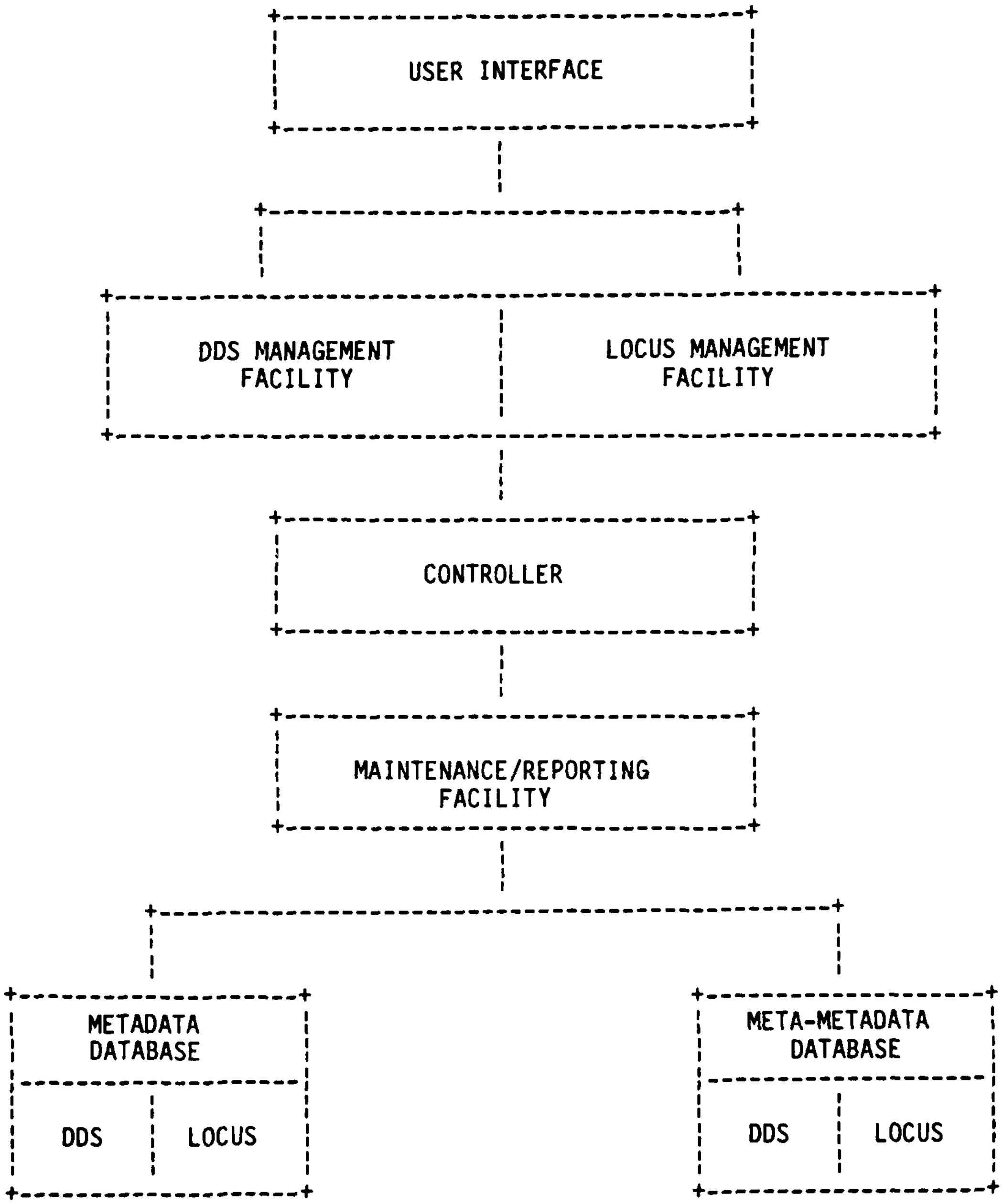


Figure 1. The LOCUS Architecture

do not conflict with each other's respective databases and schemas. The Controller may also act as a "traffic cop" between these processes in the case where maintenance functions may conflict, and may delay or prevent the execution of a process.

5.1.5 The Maintenance/Reporting Facility

The LOCUS Maintenance/Reporting Facility is the facility which issues access, update, and report requests to the Metadata Database and the Meta-Metadata Database. These requests may be directed either toward the DDS or LOCUS portion of each of these databases.

5.1.6 The Metadata Database

The Metadata Database is the database for the DDS and LOCUS. Certain descriptors in this database will be used exclusively by LOCUS, and not by the facilities in the current FIPS DDS. LOCUS will use some, but not necessarily all, of the descriptors which may reside in the dictionary.

5.1.7 The Meta-Metadata Database

The Meta-Metadata Database contains the DDS Schema and the LOCUS Schema, which define the structure of the Metadata Database, i.e., the LOCUS Database and the dictionary. There are some descriptors which may be common to both schemas.

5.1.8 The LOCUS Thesaurus Facility

The LOCUS Thesaurus Facility is not included in the architecture at this time. The precise placement of the Thesaurus Facility in the LOCUS architecture will not be decided until requirements are more precisely determined.

5.2 LOCUS/DDS Integration

Since we assume the existence of a FIPS DDS which describes the information resources of an organization, it is clear that the DDS already holds some information which, otherwise, would need to be supplied by the aspect scheme. Thus, LOCUS will encompass not only a aspect scheme for information resources, but also an interface between the aspect scheme and the DDS. This integration of LOCUS with the dictionary system will result in the following:

- The DDS contents will be expanded to include not only the metadata required to define the information environment, but also the keywords associated with the aspect scheme, and certain relationships required by LOCUS.

- This expansion of the DDS contents will require an extension to the DDS schema to support the LOCUS portion of the dictionary. In particular, the schema will be extended to support the structure for aspect trees and nodes (or, keywords) within trees. Thus, the LOCUS schema may be thought of as a subset of an extended DDS schema.

- Every descriptor which exists in the DDS schema may (or may not) be visible to LOCUS. If a descriptor is visible to LOCUS, we assume that it is so marked. It is understood that there are some logical conditions to be taken care of with this assumption, for example, if an entity-type is not visible to LOCUS, an attribute-type of the entity-type cannot be visible.

- There exists a single maintenance facility for LOCUS and the DDS. Of course, special functions may be defined that apply only to one or the other environment.

5.3 LOCUS Benefits

By combining the concepts of classification, normalized vocabulary, thesaurus, and the DDS functionality, we expect to achieve the following objectives:

- All entities will be more thoroughly classified.

- The classification of entities will be more consistent.

- Searching for candidate entities can be made more effective. Users may search with different strategies as appropriate. They may either "zoom in" or "spiral", depending on answers to specific queries.

- Searching will be done more often because the query facilities will know how to identify and deal with near misses.

- Unwanted duplication of entities defining the same concept will be reduced.

When realized, these objectives will significantly enhance the usability of current and future DDSs in the management of information resources.

6. UNRESOLVED LOCUS ISSUES

Several issues must still be addressed in the on-going design of the LOCUS functionality and architecture. These issues are:

- Where does the Thesaurus Facility fit into the LOCUS architecture?

- What are the rules and implications of assigning keywords to aspect trees?

- How does the User Interface determine whether a user request is directed toward the DDS or LOCUS?

- How does the User Interface operate?
- What is the extent of interaction between the DDS Management Facility and the LOCUS Management Facility?

The resolution of these issues will be addressed in a future report.